Cluster Based Text Mining

Nikita R. Andhrutkar, Professor Prashant M. Yawalkar

Nikita R. Andhrutkar, Dept. of Computer Engineering, MET's Institute of Engineering Nashik, Maharashtra, India
Professor P. M. Yawalkar, Dept. of Computer Engineering, MET's Institute of Engineering Nashik, Maharashtra, India

---***---

Abstract - *Discovery of relevance feature in text document is seems as an efficient approach to decide document relevance i.e. relevant or irrelevant in the domain of information retrieval. Existing techniques are based on term-based approach. Term-based approach mines the terms from training set to depict the relevant features, but it suffers from the problem of low level support. Many traditional approaches of relevance feature discovery have some issues of polysemy and synonymy in which same word in different context has multiple meanings. Though problem of term based approach resolved in pattern based text mining approach but it experienced with the large number of noise patterns. Many times it happen that search an engine retrieves the number of document in the response of user query which may irrelevant. For example, for query 'cat', user interested for 'cat entrance exam' but search engine output the 'cat' i.e pet. Therefore, there is need of an efficient tool which helps to deduct unwanted documents by extracting relevance between different documents as per user query.*

***Key Words*: Text mining, text feature extraction, text classification, Feature Discovery, Pattern Mining**

1.INTRODUCTION

A web search engine arranges the documents by the relevance to the user query. To arrange the documents web search engine required their relevance score to establish relevance of documents to the user query. Generally, information retrieval system is associated with the web search engines. In IR, documents are retrieved from the digital collections such as, news, corporate reports, abstracts etc. The method RFD (Relevance Feature Discovery) extracts the relevance rank between documents which is useful to categorized text documents into relevant or irrelevant category. But there are some challenging problems discovered in pattern based mining techniques. Low level support problem is identified in term based approach. IR mainly identifies the usual documents of unstructured nature. IR hides the problems which satisfies the core definition. In general, long patterns are more specific for topic discovery but they have limitations of low frequencies and support. Another problem is "misinterpretation". In this the support and confidence are used for text mining but they are not suitable for the problem solving approach. To addressed these problem PTM model have been proposed in [1], [4],[5]. In PTM model to find out useful features based on calculation of weight the sequential patterns mine from textual paragraphs and dispose over term space. Concept-based model i.e. CBM [2], [3] have been proposed to determine the concept using NLP techniques. To discover the concepts from sentences verb-argument structures is used but still the problem to merge the patterns into relevant and irrelevant document is alive. In recent years, multiple term-based techniques are developed for document ranking, information filtering and classification of texts [6], [7]. There are two term based models are used to discover feature terms only from relevant and unlabelled documents. Also many hybrid methodologies have been developed for text message category classification. Rocchio classifier is utilized in first phase to extracts set of original irrelevant documents from unlabelled set. In the second phase, support vector machine classifier is used to classify text documents. With these two classifier a term-based model and pattern mining get efficiently mine for information filtering systems. PTM is feasible data mining technique to the text mining area. PTM is not desired method due to its low capability of dealing with the mined patterns. PTM model for user profile filtering task in which system aims to filter out non-relevant document that incoming according to profile of user, once topic or the patterns are gained using PTM, centroid i.e. feature vector is used to handle representation of area of topic. In this extracted patterns from training set are represented. A similarity based concept is implemented to determine similarity between documents. To extracts the concept from document the system scans the documents. Their proposed concept-based mining algorithm is developed.

It is more specific to consider distribution and specialties of query terms for relevance feature discovery. With this all, it seems more difficult to extend specificity containing conditions only because feature specialty is depending upon perspectives of users. To build ranking functions 'n-gram' is more selective to carry more "semantic" rather than words. It is useful to develop good ranking functions [8],[9]. Rule-based Natural Language Processing and Context Free Grammar techniques used for phrase based text representation. It is mainly for web document management. Language model proposed to calculate weights for 'n-grams' which is approximated by unigram, bi-gram and tri-gram model with word dependency consideration.

2. RELATED WORK

In this section we are going to discuss related workdone. In this literature survey we also focus on relevance feature extraction technique proposed and utilised in previous systems as follow:

In 1999 Fei Song et al.[1] proposed new language model for retrieving information. It is based on range of data smoothing with Good-Turing estimate, curve-fitting functions, and model combinations. The proposed model is conceptual simple and intuitive. It can easily extend to organize the probabilities of phrases including pairs and triples of words. The technique developed for smoothing of data can be easily embedded into proposed language model. It referred as the general model or framework for language based information retrieval. In future work they were planning for to include a number of extensions to our language model for information retrieval.

In 2002, F. Sebastiani[2], researched about TC i.e. text categorization within information system. The categorization is nothing but classification of texts or spotting of texts. It is applied in many contexts varying from document indexing based on a controlled vocabulary, to document filtering, automated metadata generation. KE is knowledge engineering defines the set of rules on how to classify documents in specified categories. In text categorization machine learning researchers found challenging applications in which dataset consisting of thousands of documents and categorized by tens are widely used. To check that learning technique can scale up to substantial size TC is the good benchmark.

In 2004, Sheng-Tang Wu et al.[3]. Proposed two approaches based pattern deploying strategy. The have investigated their performance on Reuters dataset RCV1. They have discussed about information retrieval techniques such as, PTM i.e. Pattern Taxonomy Model. According to them PTM is feasible data mining technique to the text mining area. PTM is not desired method due to its low capability of dealing with the mined patterns. Therefore, for effective discovery of patterns they represented two algorithms such as, PDM i.e. Pattern Deploying Method and PDR i.e. Pattern Deploying with Relevance Function. In PDM, sequential patterns are deployed into feature space and the relations between patterns described as "is-a" relation in PTM. Whereas, in PDR, it uses relevance functions and utilized a probabilistic method for estimation of weight of term. SPMiner algorithm is used to retrieve a set of frequent sequential patterns. To enhance the effectiveness of pattern based method they shown pattern refinement as key improvement.

In 2006, Sheng-Tang Wu et al [4] represented pattern taxonomy extraction to extract descriptive frequent sequential patterns by deducting meaningless ones. They have used pattern based model with frequent sequential pattern instead of keyword based concept. It is mainly used to solve the problem of mining sequential patterns from text documents. PTM is tree like structure that illustrates the relationship between extracted patterns from the collection of text and it

also helps to prune meaningless patterns from pattern taxonomy. They were applied PTM for user profile filtering task in which system aims to filter out non-relevant document that incoming according to profile of user. Once topic or the patterns are gained using PTM, centroid i.e. feature vector is used to handle representation of area of topic. In this extracted patterns from training set are represented.

In 2006, S. Shehata [5], proposed concept based analysis of text features and concept based similarity measures. They discussed about the concept of clustering which is traditional approach to club similar types of documents into cluster. Text clustering approach, documents are spread into different clusters according relevant topics of each document. Clustering of documents represents the classification and clustering of documents. To find important terms from text documents term frequency is calculated. A similarity based concept is implemented to determine similarity between documents. To extracts the concept from document the system scans the documents. Their proposed concept-based mining algorithm is developed using sentence-based concept analysis, document-based concept analysis, the corpus-based concept analysis and the concept-based similarity measure. Better quality is achieved in the process of clustering by exploiting the semantic structure of the sentences in documents. And the proposed concept based algorithm performance is better than the existing single term based approaches.

In 2007, Shady Shehata et al[6]. represented concept-based model. It analyzed the terms on sentences as well as document level than analysis of documents. Proposed model contains concept-based statistical analyzer, conceptual ontological graph representation and concept extractor. In this model, each sentence in the text document is labeled automatically with the help of PropBank notations. Both verb and arguments are considered as terms. In sentence, there may have one argument but more than one verb. Model also consists of concept-based statistical analyzer, representation of COG and concept extractor. To maintain sentence semantics concept based statistical analyzer is used. Weighted based on its position is the COG representation and to merge the weights computed by concept-based statistical analyzer concept extractor is utilized.

In 2007, Donald Metzler et al [7], proposed a robust query expansion technique. It is used for information retrieval based on Markov random model. The proposed technique is also referred as,"Latent concept expansion". At the time of expansion, it provides the mechanism for term dependencies. They have evaluated technique against the relevance model. For multi term expansion LCE is utilized to perform single or multi-term expansion. The technique produced high quality, well construct and topically relevant multi-term expansion concepts. Lastly, they discussed that LCE is capable of capturing syntactic and query-side semantic dependencies.

In 2008 Georgiana Ifrim [8], accomplish the task of weaken a-priori required knowledge about database as well as tokenization result with the character length. Gradient ascent is used for accomplishment of the task in the space of all 'n'-grams. They discussed about bag of word representations used for categorization of text. A typical type of pre-processing such as, stemming or removal of words is used to provide training to the text. But it required detailed knowledge of text language for categorization. Their contribution is for Structured Logistic Regression i.e. SLR which incorporates the best features of 'n'-grams for variable length. To maximizing the logistic regression likelihood of the training data they have developed a coordinate-wise gradient ascent technique. It inherits the structure of n-gram feature space

In 2009, C. D. Manning [9] did the study of information retrieval system. IR mainly identifies the usual documents of unstructured nature. IR hides the problems which satisfies the core definition. Generally, unstructured data consists of the data that is not clear but the fact is no data is unstructured. IR covers the supporting users in filtering collection of documents. The proposed task is similar to arrange or sort the books into book-shelf as per topic.

In 2015 Yuefeng Li,et al., .[10] discussed about the problem of existing text mining and text classification techniques. All are adopted term-based approaches. They analyze that the previous techniques suffered from the problems of polysemy and synonymy. Also they demonstrate that effective tools are required to effectively use large scale patterns. They have proposed relevance feature discovery (RFD) to find relevance features present in the text documents. They addressed two challenging issues in text mining such as, low-level support and pattern mining. Continued with RFD model they have implemented WFeatures and FClustering algorithms. FClustering algorithm describes the feature clustering process and discovers the set of patterns whereas; WFeature algorithm is used for computations of weight of classified terms.

3. SYSTEM ARCHITECTURE

Fig 1. represents the proposed system architecture. It is divided into two phases such as:

1. Document Classification

2. Pattern Categorisation

Proposed technique for relevance feature discovery is innovative model for relevance feature discovery. It discovers both positive and negative patterns in text documents as higher level features and deploys them over low-level features (terms). It also classifies terms into categories and updates term weights based on their specificity and their distributions in patterns. The proposed FClustering is Feature Clustering algorithm discovers both positive and negative patterns in text documents as higher level features and deploys them over low- level features (terms). It also classifies terms into categories and updates term weights based on their specificity and their distributions in patterns.

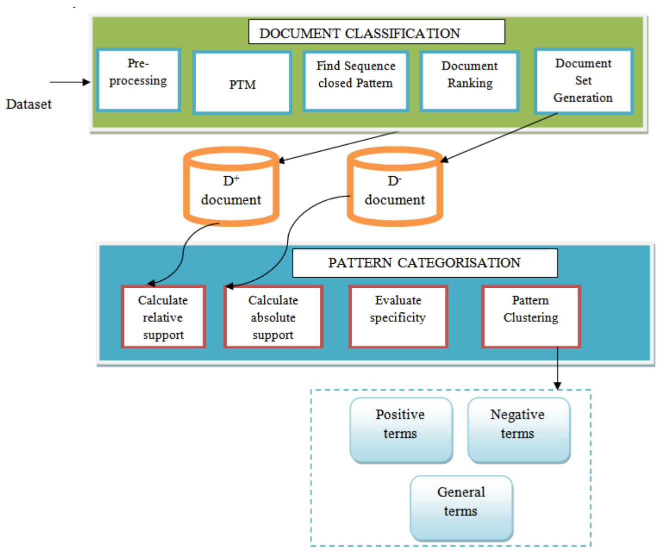

Fig 1: System Architecture

4. CONCLUSIONS

In this review paper we have studied some existing techniques of text mining. Text mining is the process of deriving high quality information from given text. There are several existing techniques available for text mining which is based on term based approach such as, PTM, PDR, LCE, PDM etc. But they suffered from some challenging issues such as, low level support, polysemy and synonymy in which same word in different context has multiple meanings, large number of noisy patterns. From literature survey analysis, we analyzed that there is need of such technique which discovers the positive and negative patterns from the text documents. Lastly, we conclude that RFD is relevance feature discovery technique seems to be better solution to extracts the relationship between textual documents.

REFERENCES

[1] F. Song and W. B. Croft, "A general language model for information retrieval," in Proc. ACM Conf. Inf. Knowl. Manage., 1999, pp. 316–321

[2] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surveys, vol. 34, no. 1, pp. 1–47, 2002

[3] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic pattern taxonomy extraction for web mining," in Proc. Int. Conf. Web Intell., 2004, pp. 242–248.

[4] S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in Proc. IEEE Conf. Data Mining, 2006, pp. 1157–1161

[5] S. Shehata, F. Karray, and M. Kamel, "Enhancing text clustering using concept-based mining model," in Proc. 2nd IEEE Conf. Data Mining, 2006, pp. 1043–1048

[6] S. Shehata, F. Karray, and M. Kamel, "A concept-based model for enhancing text categorization," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2007, pp. 629–637.

[7] D. Metzler and W. B. Croft, "Latent concept expansion using Markov random fields," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007, pp. 311–318.

[8] G. Ifrim, G. Bakir, and G. Weikum, "Fast logistic regression for text categorization with variable-length n-grams," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2008, pp. 354–362.

[9] C. D. Manning, P. Raghavan, and H. Schtze. Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[10] Y. Li, A. Algarni, M. Albathan, Y. Shen, and M.A. Bijaksana, "Relevance Feature Discovery for Text Mining," in IEEE Trans. Knowl. Data Eng., vol. 26, no. 6, pp., Jan. 2015.

BIOGRAPHIES

Andhrutkar Nikita R. has completed Bachelor's Degree from Late PVG College of Engg. Nashik. Now pursuing ME Degree in Computer Engineering from MET's Institute of Engineering, BKC, Nashik.

Prof. Yawalkar Prashant M. is working in MET's Institute of Engineering, BKC Nashik. He has published/presented research papers in National and International Journals/Conferences. His area of interest is in Image Processing and Sub computing.